Effective Use of Big Data Planning and Analytics in Crop Planting

T. Karthick
Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu. E-mail: karthict@srmist.edu.in

T. Kirthiga Devi*
Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu. E-mail: kirthigt@srmist.edu.in

M. Ramprasath
Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu. E-mail: mramprasath@gmail.com

Abstract

The effective rainfall, temperature and humidity received in a particular region plays a huge role in the development of crops in agriculture during all time lapses of yield. There could be instances when the constraining, precipitation, temperature and humidity probably won't be adequate to help crop development and along these lines information about it already can help crop growers in assessing the kind of crop that needs to be grown for maximum profit considering the farmer's contribution. Expectation of viable precipitation, temperature and humidity needs is a difficult chore which requires an accurate and conscientious investigation of a compelling synopsis of components, for instance, humidity and temperature. A few years ago successful rainfall was registered by contemplating triple central points, the humidity, climate and the resulting precipitation. Various numerical models have been planned and expected over many years and a considerable amount of it are still applied to locate the 'successful precipitation'. Here the forceful precipitation, temperature and humidity is registered by an advancement of difficult mathematical calculations which have been divided into more straightforward approaches and processes and given as an expected outcome. In the expected work, we foresee reasonable earning of crops for development dependent on atmospheric characteristics such as precipitation, temperature and humidity.

Keywords

Crop Planning, Data Analytics, Weather Regulation, Yield, Farming Development, Navie Bayes.

Introduction

Weather regulating and anticipating measures of precipitation, temperature and humidity that will happen in a zone can be dreary. It would include cohesive perception of climatic circumstances and cloud placement along with the management of 'models' to imitate environmental circumstances which also 'prompts' a high level of multifaceted and varied design. Additionally, estimating the measure of precipitation that might be viable is burdensome. Estimating the effective capacity of precipitation involves testing the dust attributes, ground attributes, 'groundwater' quality and many aspects to calculate an accurate evaluation. The ultimate goal of finding the suitable crop for farmers to grow this technique considers the attributes which are extremely critical with respect to precipitation [1]. The amount of water exploited for the 'water' system is anticipated without contemplating the successful measure of precipitation that will be known about a specific zone. This might cause wrong judgements about measures of 'water' to be used.

Ultimately when a great deal of 'water' is employed it may prompt overuse of 'water' systems, water clogging and might bring about saltiness in this way declining harvest yield. Also, usage of restricted 'water' might give an outcome of an 'under water' system and diminished output of crop effectiveness. In this manner breaking down the right quantity of 'water' that is provided for flooding the crops is critical. Strategies include testing distinctive geological attributes, for instance, area geography, inclination, soil surface, structure and profundity, 'mean temperature', normal precipitation, humidity to look over the assess of precipitation that might be obtained indicated locations. Utilizing it, the harvest 'water' prerequisites is solved all the time. The nonexclusive advances included appear in the chart [2].

When the precipitation has been measured and processed, utilizing information mining methods it is utilized to foresee the everyday 'water' prerequisites of the produce with an aim that ranchers have any information in drawing a rough approximation of the right quantity of 'water' needed for the water system. The aim of the technique is to advise ranchers in using ideal yields for advancement and advise them to choose the reasonable produce for the ideal advancement that is dependent on given characteristics. Information scooping strategies, for instance, are used to carry information about the quantity of 'water' necessary for a water system who would reinforce the cultivation[3]. This additionally causes ranchers to know the produce reasonably in the land in their district.

The target of our experiment is to give an answer for 'Smart Agriculture' by observing horticultural fields which help the producer in expanding profitability as it were. Climate

figure information acquired from IMD (Indian Meteorological Department, for example, mean temperature, precipitation and normal humidity storehouse gives knowledge into which harvests are reasonable to be developed in a given territory. In this way, the proposed framework takes the topographical area of the client as an info. From the area, the land dampness is acquired. The preparing part likewise has two additional informational indexes for example one got from the climate calculation division, evaluating the weather conditions anticipated in the ongoing year and different information being fixed.

This fixed information is yield creation and information alluding to requests of different types of harvests acquired from a few government sites. The proposed framework applies Machine Learning and forecast calculations like 'Decision Tree', 'Naive Bayes' and 'Random Forest' algorithms to distinguish the example among information and afterward process it in accordance to enter circumstances [4]. This thus will suggest the most favorable and reasonable harvests as indicated by the given ecological elements. Subsequently, this framework will just require the land area of the client and it will propose the quantity of beneficial yields giving a decision legitimately to the farmer about which harvest to develop. As earlier year's creation is additionally considered, the forecast will progressively produce exact results.

India being a tremendous nation that exclusively relies upon farming, its economy overwhelmingly relies upon agribusiness yield development and agro mechanical items. In India, farming is basically impacted by water which is entirely eccentric. Farming development likewise relies upon differing Soil dampness, Surface temperature and furthermore on climate angles which incorporate temperature, precipitation, humidity and so forth. As of now, India is quickly advancing towards a specialized turn of events[5]. Consequently, innovation will be valuable to agribusiness which can expand crop efficiency prompting better respect for the rancher. Agribusiness is the broadest financial area and assumes a significant job in general improvement of the nation. Around 60 % of the land in the nation is utilized for farming so as to get the job done the requirements of 1.2 billion individuals.

Accordingly, modernization of agribusiness is quite significant and along these lines will lead the farmers of India towards benefit. Decision Analysis (DA) is the technique for investigating informational collections to represent derivations of the report of information they consist of, progressively with the guide of particular frameworks and programming. Prior produce expectation was achieved by thinking about the rancher's understanding on a distinct area of land and harvest. Be that as it may, as the

circumstances adjust progressively yet quickly, producers are compelled for developing an ever increasing number of yields. This being the present circumstance, the vast majority of them need more information about the new yields and are not totally mindful of the advantages they get while cultivating them. Likewise, the harvest profitability can be improved by comprehension and anticipating crop execution in an assortment of natural conditions[6].

A target technique to figure precipitation inclusion with a neural system is introduced. This procedure utilizes the whole information accessible as indicators at neighborhood stations including both numerical model outcomes and climate information obtained after the model starting time, which once in a while repudiate one another and henceforth must be taken care of cautiously by all around experienced forecasters. Since the technique gives a veritable and practical gauge of areal precipitation inclusion, its expertise scores are better than those of the diligence figure (after 3 h), the direct relapse conjectures, and numerical precipitation expectation model. This ANN model is basically utilizing radar, satellites and climate station information joined with numerical items given by the Japan Meteorological Agency (JMA). As the ANN utilized just 1 year information for preparing, the outcomes were confined[7].

The precipitation is one of the most huge informational collections of water assets on the board. Utilizing the month to month precipitation information in the years extending from 1941 to 1999 from 245 precipitation screen stations close to Chao Phraya River in Thailand, the precipitation estimation with a counterfeit astute strategy is sensible. Counterfeit neural systems is one the most broadly directed strategies of information mining. It tends to be applied on prescient mining undertakings to make a visualization. The primary commitment of this paper is to utilize a neural system model for month to month precipitation forecast. The preparation and testing designs are created as a period arrangement information of the previous ten months. The determined quantities of preparing and testing designs are 372 and 96, individually. In the preparation procedure, the neural system gives 99.6 % exactness and 96.9 % of precision in the testing procedure[8]. The outcomes demonstrate that it is conceivable to extend yearly precipitation one year ahead with acceptably exactness.

The investigation supported that it may be smarter to embrace different related climatic factors, for example, wind speed, darkness, surface temperature and pneumatic force as the extra indicators. The benefits of utilizing counterfeit neural system methods to clarify the nonlinear conduct between the info and yield is investigated to evaluate the precipitation in rainstorm. The model uses the information of going before long stretches

of precipitation information in storms just to conjecture the rainstorm precipitation of the coming year. Month to month precipitation time arrangement information is utilized for the present investigation[9]. This model records the info yield non-straight relations and anticipated the storm precipitation precisely during the autonomous period very well. All India rainstorm precipitation estimates were created by utilizing territory weighted precipitation gauges of all the sub-divisions. Just because up-scaling is advanced in storm precipitation expectation utilizing neural system procedure in catching the inconstancy of the all India precipitation better. Back spread ANN to gauge the mean summer-storm precipitation over the entire nation. The use of this model is constrained to month to month precipitation information as it were.

The fog based Architecture was proposed by Karthick,[10] can be utilized for data alnysis. The present examination researches the capacity of fluffy principles or rationale in displaying precipitation for Nigeria in the south western locale. The definitely evolved Fuzzy Logic model is fundamentally composed of two utilitarian components; the fluffy thinking or dynamic unit and the information base. Two tasks have been performed on the Fuzzy Logic model; the fuzzification and defuzzification activity. The model evaluated yields were contrasted and the genuine constant precipitation information. Recreation results uncover that evaluated outcomes are in acceptable concurrence with determined information[11]. Expectation Error, MAE or (Mean Absolute Error), RMSE or (Root Mean Square Error) and accuracy of the prediction were determined, and based on the outcomes obtained, it very well may be recommended that fluffy technique is productively equipped for overseeing dispersed information. This created fluffy standard based model shows flexibility and furthermore the capacity in making a not well characterized connection among info and yield factors. Fluffy rationale utilized for precipitation expectation. The fluffy rationale technique is utilized to display and gauge neighborhood precipitation information. The RMSE among model and information yield is seen to be 319.0 mm which is lesser than that by utilizing either the Niño 3.4 alone of 349.2 and 1557.3 mm, individually or the neighborhood downpour[12].

India is perhaps the most seasoned nation which remains to put resources into agribusiness. Yet, as of late the patterns in this field have seriously advanced gratitude to globalization. Numerous new advancements are developing to recover wellbeing. One such system is precision horticulture. Definite agribusiness is the innovation of "site-explicit" cultivating. it's given us the benefit of adequate info, yield and better choices in regards to cultivating[13]. Despite the fact that Precision Horticulture has conveyed better enhancements despite everything confronting certain problems. There exists numerous frameworks proposing the contributions for a specific cultivating plot. Suggesting yields

is a significant area of exactness farming. Proposal of yields relies upon different parameters. Precision horticulture points in recognizing these parameters during a site-explicit way in order to determine problems with respect to trim determination. The "site-explicit" procedure has improved the outcomes yet there's a prerequisite to manage the consequences of such frameworks. There ought to be a solid expectation framework to be suggested by thinking about the traits[14].

This model predicts crops upheld precipitation, humidity and dampness. The data is pre-handled and changed over to numeric values before preparing and expectation. The three strategies applied for expectation specifically Decision Tree, Random Forest and Naive Bayes[15].

Advantages

- Give great precision for the considered dataset.
- Crop expectation exactness causes ranchers to encourage better development and yield.
Analysis and Methods:

Data Assortment

The information assortment method includes the decision of value information for checking and review. Here we utilized a dataset with four highlights to be specific soil dampness, ph, temperature, moistness and precipitation information. Crafted by a factual investigator to find ways and thorough information, clarifying it, and breaking down outcomes with the help of strategies that can be measured.

1) Data Perception

A lot of information spoken to in a structure that seems real and is more straightforward to comprehend and break down as it is exact. A few organizations determine that an examiner must have aptitudes set to make slides, outlines, diagrams and draw bits of knowledge from them while drawing out a coherent story. In our methodology, the information histogram and dissipate lattice are appeared as information representation parts.
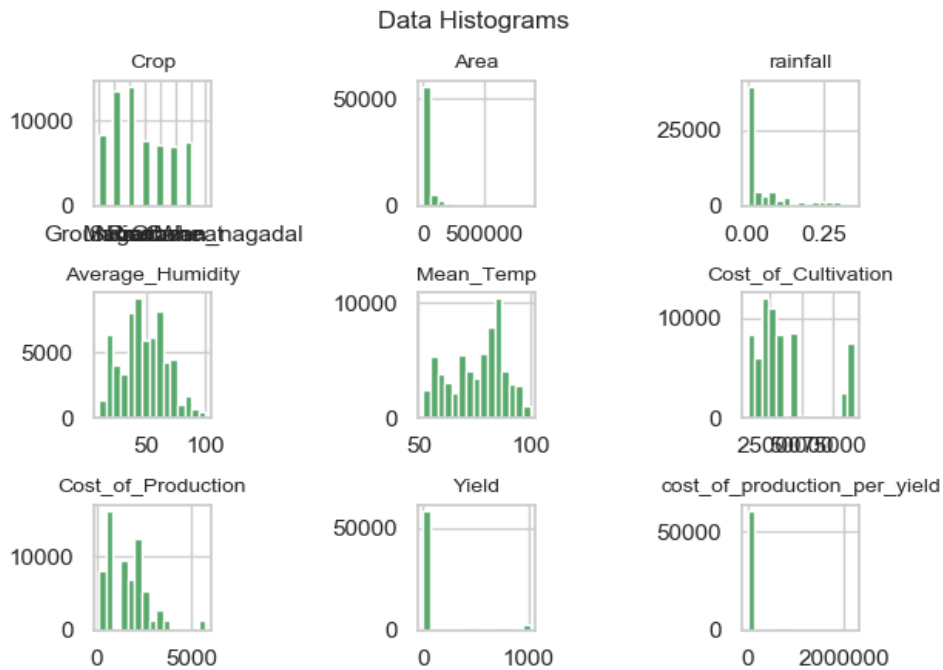
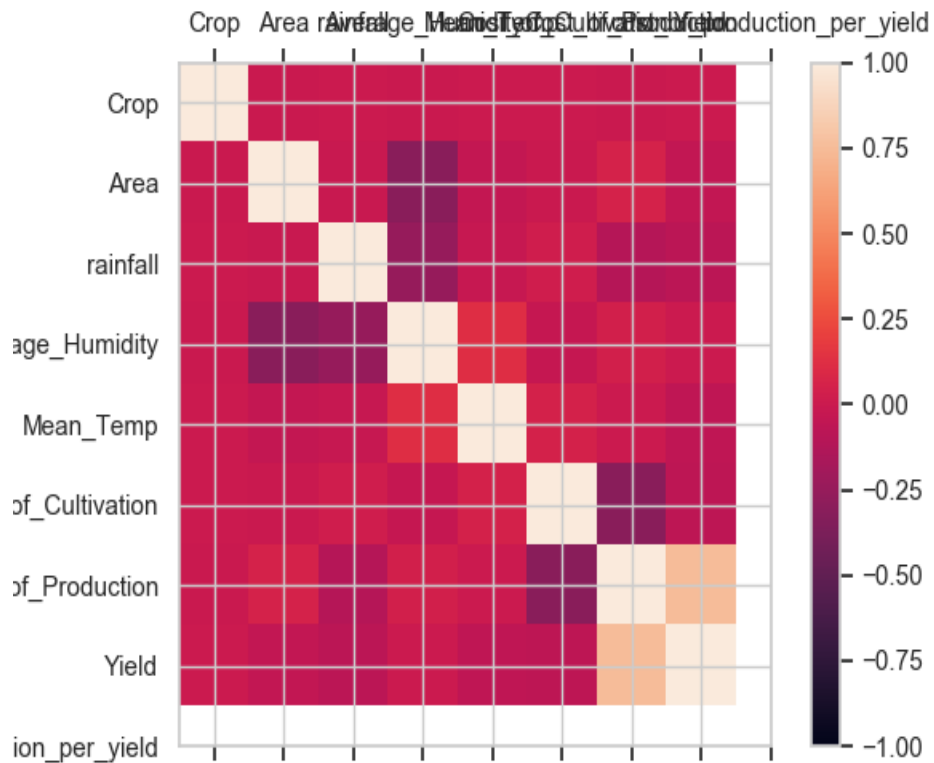**Figure 1 Graphical Representation of Dataset**

Figure 2 Graphical Representation of Dataset- Correlation plot

2) Data Preprocessing

Data preprocessing is done to change data that fits AI groups from crude information. Arranged information permits a researcher to get information in a precisely progressive way by requesting outcomes from an AI framework. The system includes data organizing, cleaning, and examining.

3) Dataset Parting

A set of data utilized for ML processing is isolated into three subsets — preparing, testing and approval set.

Training set - A training set is utilized by the algorithm to create a model of it's own by processing the data. This data set will help the algorithm to learn and predict further.

Test set - This set is needed for checking the set generated after the training and its probability for speculation and accuracy. The previous method is a set's capability to recognize designs in vague information after being prepared. It's critical to utilize various sets for preparing and testing to get rid of overfitting models, that is also the insufficiency for speculation we referenced previously.

4) Model Preparing

Once the preprocessed information is ready, the information that has been gathered and part it into training and test, the data scientist can proceed with model preparation. This procedure incorporates "Calculating prepared information. This will help process data and generate a model that is prepared to figure out an objective worth (quality) in the new data present and give a solution you would want to proceed with predictive investigation. The motivation behind preparing a model is to set up that model.

5) Model Assessment and Testing

Purpose of this step is to prepare the main model to deliver an incentive with speed and fairness. A researcher can arrive at this purpose through tuning. It is the advancement of a model structure to get a calculation's best output.

System Design

This part gives a review of the engineering plan, dataset for execution, calculations utilized and UML structures.
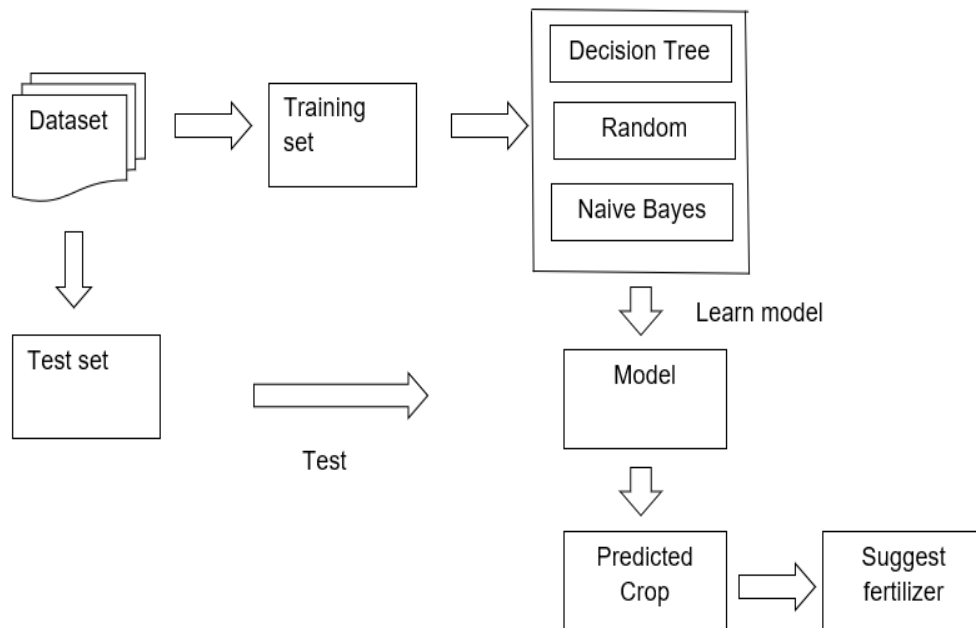
Figure 3 System Architecture

The figure above speaks to framework design of the proposed framework. First the dataset is stacked, pre-handled and partitioned and prepared into test sets then we apply ML algorithms, for example, Decision Tree, Random Forest and Naive Bayes. The anticipated outcomes are contrasted and unique qualities and we at that point assess the exhibition of accomplished calculations.

It is a normalized displaying language that empowers designers to represent, imagine, build and report the entire framework of the project. By using these diagrams to represent, it makes these relics upgradeable, easily understandable, safe and lively in execution. In object-situated programming advancement, UML is a significant angle. It utilizes pictorial format so that it can be understood by everyone easily within a glance.

Dataset Collection

The dataset used for the implementation has attributes such as state name, district name, crop name, crop year, crop season, area, temperature, humidity, rainfall, cost of production, cost of cultivation and cost of production per yield. The dataset is initially cleaned so as to get rid of the unnecessary data to get the crop suggestion based on the attribute conditions.

Database Description

STATE NAME - Denotes the state where the crop has been cultivated. (String)
DISTRICT NAME - Denotes the district where the crops have been cultivated. (String)
CROP YEAR - Denotes the year in which the crop was cultivated. (Integer)
SEASON - There are two crop seasons Kharif and Rabi. This field denotes in which crop season the crop was cultivated. (String)
CROP - This field mentions which crop was grown. (String)
AREA (hectares) - Denotes the area used to cultivate the crop. (Float)
RAINFALL (mm) - Denotes the amount of rainfall received during the time of cultivation. (Float)
AVERAGE HUMIDITY (%) - Denotes the average humidity during the time of cultivation. (Integer)
MEAN TEMPERATURE (F) - Denotes the average temperature during the time of cultivation. (Integer)
COST OF CULTIVATION - Denotes the cost incurred in cultivating one hectare of the crop. (Float)
COST OF PRODUCTION - Denotes the cost incurred while cultivating the crop. (Float)
YIELD (quintal/hectare) - The yield is a measurement of the amount of crop grown, per unit area of land. (Float)
COST OF PRODUCTION PER YIELD - Denotes the cost incurred while production per yield. (Float)

1) Decision Tree Algorithm

Decision tree is one of the most commonly used machine learning algorithms that is generally utilized in issues identified with arrangement. The decision tree is built from top to bottom. The most influential characteristic is chosen in the beginning to be made as the starting node of the tree. This attribute is chosen using various methods such as information gain, gini index, etc. The same method of finding the most influential attributes is done till the entire decision tree is built.

Decision Tree follows these steps to perform the prediction,

- Spot the most influential trait of the dataset by using various methods and make it as the root node of the tree.
- Now excluding that attribute, pick the most influential attribute out of the other attributes.
- Rehash the initial 2 stages on every subset until all the attributes are placed in the most accurate order.
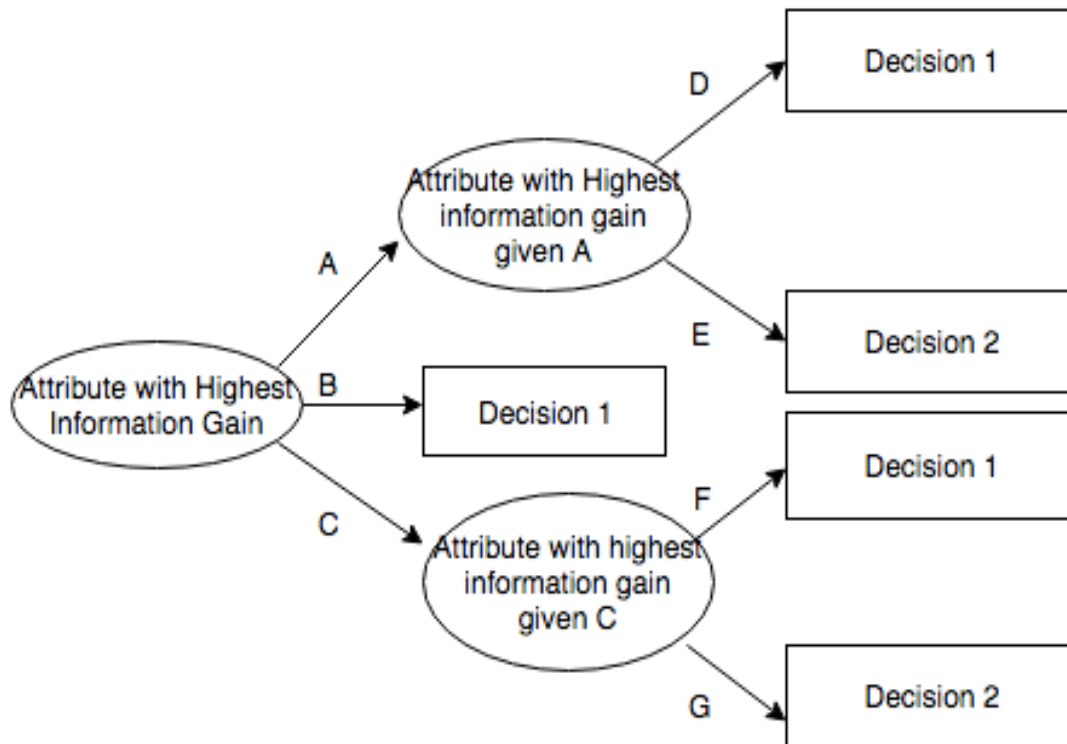
Figure 4 Flow chart Decision Tree algorithm

2) Random Forest

Random forest algorithm is nothing but an enhanced version of the decision tree algorithm. Multiple decision trees are combined to form a forest. Now, as these trees are built randomly by not using any of the methods available, these are known as random trees. So, these random trees are combined to form a random forest.

1. Create a number of decision trees in a random way without using any methods that show the most influential attributes.
2. Now, input the new data for which the prediction has to be carried out into all the random decision trees created.
3. The result that majority of the decision trees give will be the most accurate prediction.
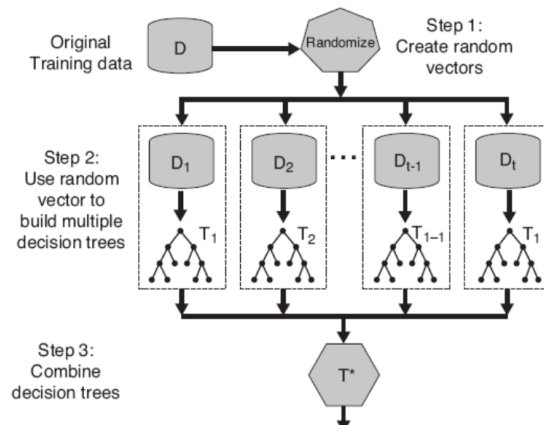
Figure 5 Flow chart of Random Forest

3) Naive Bayes

Naive Bayes classifier is a mix of different characterization calculations dependent on Bayes' Theorem. It is a group of calculations and not one single calculation, where every one of them share a typical rule, for example each pair of traits being ordered isn't reliant on one another.
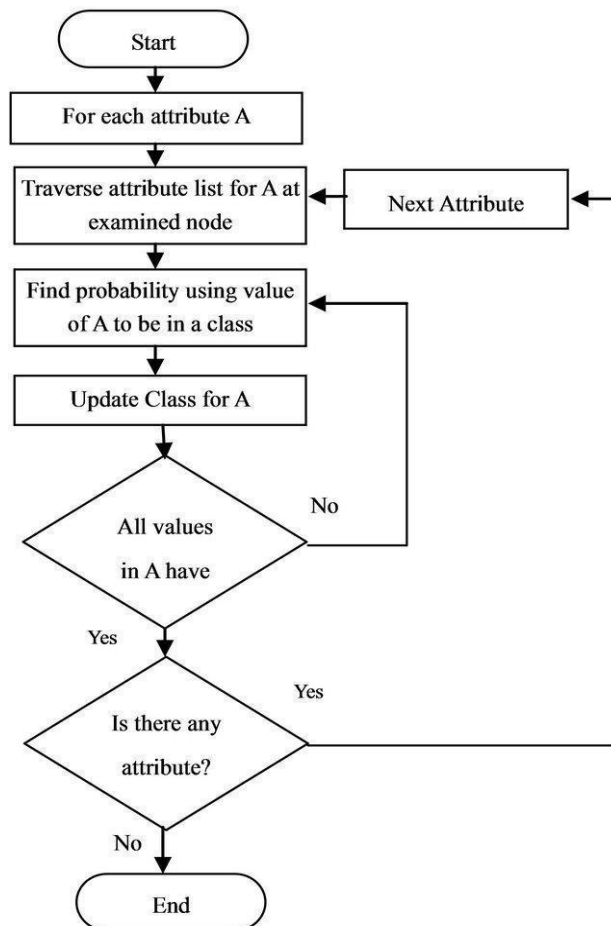

Figure 6 Flowchart of Naïve Bayes

This is for the most part utilized for record characterization issues, i.e whether an archive has a place with the classification of sports, legislative issues, innovation and so on. The recurrence of the words present in the report are utilized as the traits by the classifier.

Result and Performance Analysis

Here, Python 3.6.5 was used for the proposed idea. It was implemented using libraries such as scikit-learn, matplotlib and panda. The agriculture dataset is applied to ML calculations, for example, Decision Tree, Random Forest and Naive Bayes. We utilized these ML algorithms and distinguished the best crop to be cultivated. The outcome shows that crop expectation is effective utilizing the Random Forest calculation. Irregular Forest accomplishes 75% exactness, while Decision Tree and Naive Bayes accomplish 70% and 65% precision individually. The accompanying table shows the precision accomplished in our test study.

Table 1 Experimental Results of proposed system

Algorithm	Accuracy (%)
Decision Tree	70%
Random Forest	75%
Naive Bayes	65%

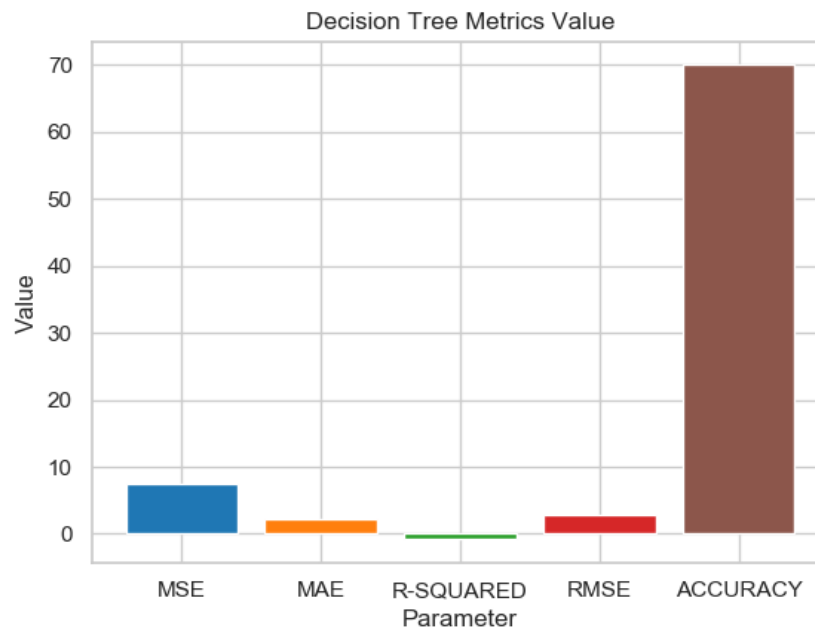The below figure shows the evaluation metric of the Decision Tree algorithm.

Figure 7 Evaluation of Decision Tree algorithm

The below figure shows the evaluation metric of Random Forest algorithm.
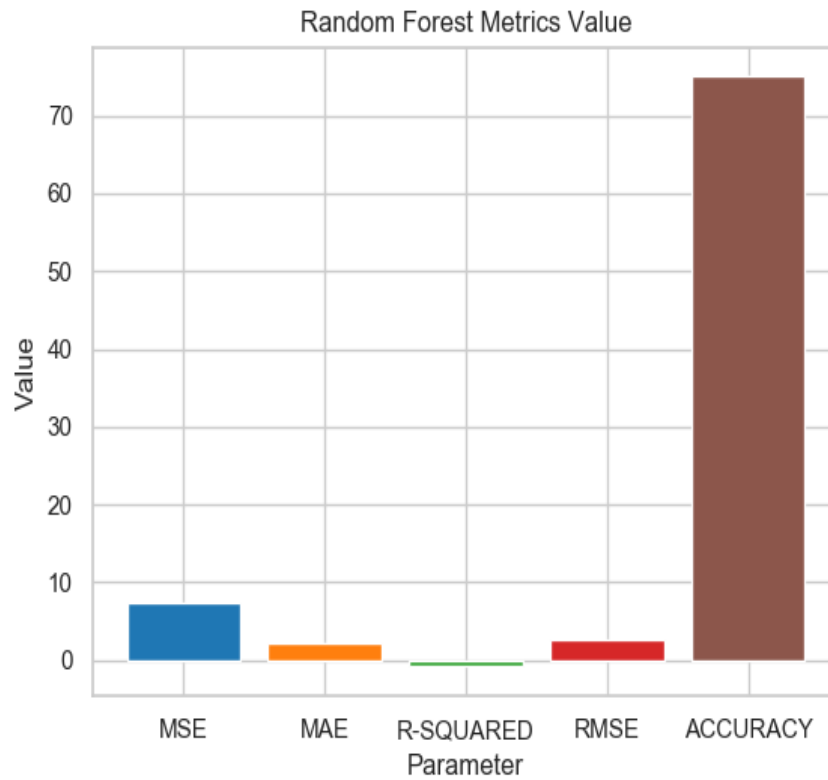

Figure 8 Evaluation of Decision Tree algorithm

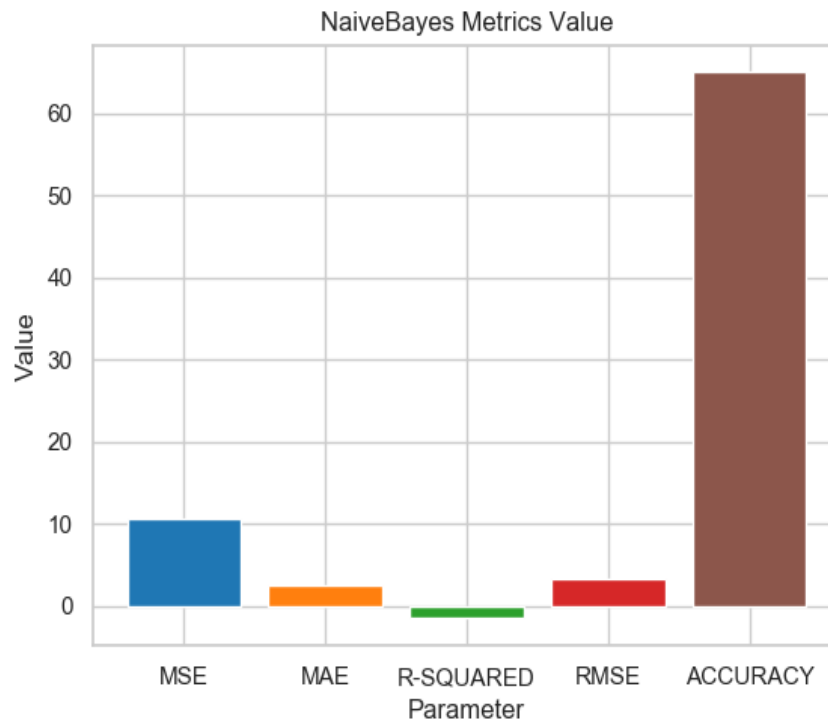The below figure shows the evaluation metric of Naive Bayes.

Figure 9 Evaluation of Naive Bayes algorithm

Error Evaluation

1) Root Mean Square (RMSE)

"Root mean square error" (RMSE) is an every now and then utilized proportion of the distinction or deduction between values anticipated by a model and the outcomes are recognized.

2) R Squared Value

"R squared" is a segment and investigative proportion of how immediate or close the information is to the fitted relapse line. It's otherwise called the cooperative of different assurances for various relapses.

3) Mean Absolute Error (MAE)

It is the aggregate of the total contrasts among expectations and genuine qualities. It gives a thought of how wrong the counts and expectations are. The measures and results gives a thought of the greatness of the error, but no thought of the heading.

4) The Mean Squared Error (MSE)

It is almost similar to the MAE which gives a good idea of the level of inaccuracy. The screenshots below speaks to exactness and conviction of the machine learning algorithms Decision tree, Random forest and Naive Bayes for phishing site discovery.


```
MSE VALUE FOR Decision Tree IS 7.628083
MAE VALUE FOR Decision Tree IS 2.165917
R-SQUARED VALUE FOR Decision Tree IS -0.861726
RMSE VALUE FOR Decision Tree IS 2.761898
ACCURACY VALUE Decision Tree IS 70.000000
```

Figure 10: Decision tree values


```
MSE VALUE FOR Random Forest IS 7.508753
MAE VALUE FOR Random Forest IS 2.139004
R-SQUARED VALUE FOR Random Forest IS -0.821035
RMSE VALUE FOR Random Forest IS 2.740210
ACCURACY VALUE Random Forest IS 75.000000
```

Figure 11: Random forest values

Figure 12: Naïve bayes values

Conclusion

India is probably the most established nation which remains rehearsing farming. Be that as it may, as of late the patterns in agriculture have definitely advanced gratitude to expansion and improvement. Numerous new innovations are developing to recover wellbeing and development of the agriculture sector in India. One such strategy is accuracy farming. The strategy was made with the plan to conquer 3 hindrances. Firstly we have to locate the successful measure of precipitation, temperature and humidity, also we have to utilize the compelling precipitation to discover the water system water required and in conclusion is to advocate reasonable yields that ought to be appeared by the farmers to build crop efficiency. The proposed model anticipated harvest using the atmospheric characteristics such as precipitation, temperature and humidity.. Dataset is cleansed and changed over to numerical values before preparing and forecast. We have used three strategies or algorithms for optimal crop prediction namely Decision Tree, Random Forest and Naive Bayes. The strategy is fruitful in every one of the three angles and in the ensuing years it is relied upon to bring more data of various regions as we know that more accurate data about the same is required to improve the model to make it more accurate.

References

Koizumi, K., & Weather, F. (2012). An objective method to modify numeric model forecasts with newly given weather data using an artificial neural network. *IEEE Control and System Graduate Research Colloquium, ICSGRC*, 82-87.

Manusthiparom, C., Oki, T., & Kanae. (2003). Quantitative Rainfall Prediction in the country of Thailand. *First International Conference on Hydrology and Water Resources on Asia Pacific Region (APHW)*.

Guhathakurta, P. (2006). Long-range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network. *Current Science*, *90*(6), 773-779.

Halide, H., & Ridd, P. (2002). Modeling inter-annual variation of a local rainfall data using a fuzzy logic. *Proceedings of International Forum on Climate Prediction, James Cook University, Australia*, 166-170.

Nishit, J., Amit, K., Sahil, G., Vishal, P., & Prajakta, K. (2018). Crop Selection Method on the basis of Various Environmental Factors Using Machine Learning. *International Journal of Advanced Research in Computer and Communication Engineering*, *7*(10), 103-106.

Dahikar, S.S., & Rode, S.V. (2014). Agricultural crop yield prediction using artificial neural network approach. *International journal of innovative research in electrical, electronics, instrumentation and control engineering*, *2*(1), 683-686.

Kumar, R., Singh, M.P., Kumar, P., & Singh, J.P. (2015). Crop Selection Method to maximize crop yield rate using machine learning technique. *In international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)*, 138-145.

Karthick, T., Suresh, S.S.P., Prathyusha, V., & Niharika, A.V. (2017). Intelligent IoT Based Water Quality Monitoring System. *International Journal of Applied Engineering Research*, *12*(6), 5447-5454.

Anandan, M., Manikandan, M., & Karthick, T. (2020). Advanced Indoor and Outdoor Navigation System for Blind People Using Raspberry-Pi. *Journal of Internet Technology*, *21*(1), 183-195.

Karthick, T., & Manikandan, M. (2019). Fog assisted IoT based medical cyber system for cardiovascular diseases affected patients. *Concurrency and Computation: Practice and Experience*, *31*(12), e4861. https://doi.org/10.1002/cpe.4861

Shakoor, M.T., Rahman, K., Rayta, S.N., & Chakrabarty, A. (2017). Agricultural production output prediction using supervised machine learning techniques. *In IEEE 1st International Conference on Next Generation Computing Applications (NextComp)*, 182-187.

Veenadhari, S., Misra, B., & Singh, C.D. (2014). Machine learning approach for forecasting crop yield based on climatic parameters. *In IEEE International Conference on Computer Communication and Informatics*, 1-5.

Jain, N., Kumar, A., Garud, S., Pradhan, V., & Kulkarni, P. (2017). Crop selection method based on various environmental factors using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *4*(2), 1530 -1533.

Igor, O., Renato, L.F.C., Bruno, S., & Marco, A.S.N. (2018). A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast. *14th IEEE eScience*. https://arxiv.org/abs/1806.09244

Rao, R.N., & Sridhar, B. (2018). IoT based smart crop-field monitoring and automation irrigation system. *In IEEE 2nd International Conference on Inventive Systems and Control (ICISC)*, 478-483.